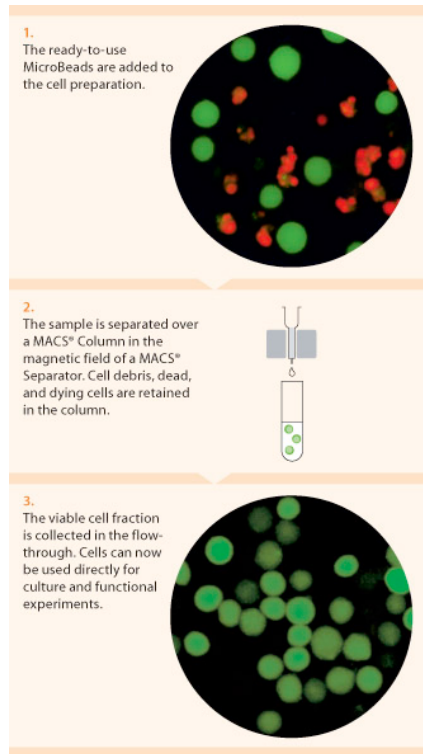
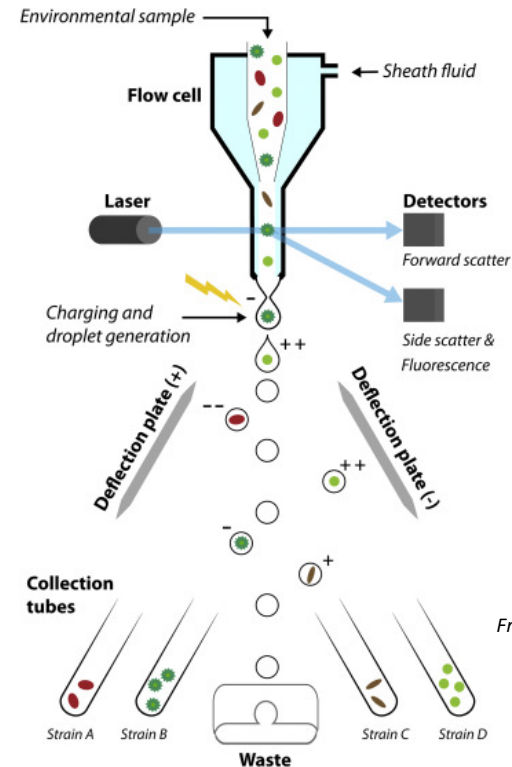


To achieve the cell viability of cryopreserved cells suitable for 10X 3' sequencing, would it be worth comparing commercial dead cell removal kit vs. FACS sorting directly into 10X library for removing dead cells after thawing?



From Miletnyi

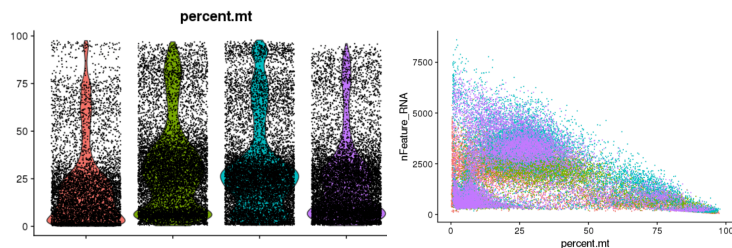
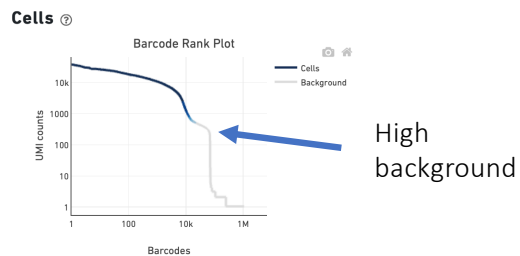
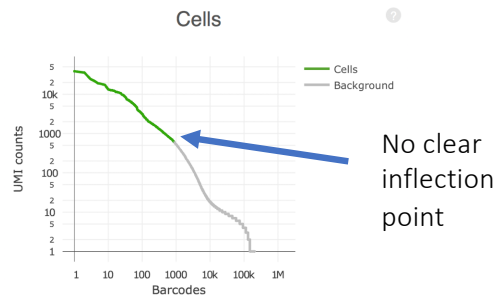


From Pereira et al 2018

- Can be more 'gentle' on sample
- Usually not 100% pure
- Easy for any lab to implement

- Higher pressure & speed = more cell damage
- Low pressure sorters exist (throughput vs gentle)
- Higher purity at sort point
- Access and scheduling

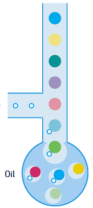
We are planning of a pilot experiment of 10X 3' RNA for a couple of 8-plexed reaction before we expand onto larger samples. In addition to the standard QC metrics (e.g. viability, number of detected cells and genes per cell, mitochondrial content etc) what would you recommend we pay attention to determine the quality of the data?



Indication	Suggests	Strategy	Keep in Mind
Poor or multiple inflection points	May have mixture of 'good' cells and lower viability or contaminating cells (like RBCs)	Sample prep to improve viability and/or enrich for higher viability cells	Any purification or enrichment can also introduce some biases – what was the low viability cell type removed?
High ambient background; low fraction of reads in cells	Dead cells or debris remained in the cell suspension	Additional washes or purification to remove, if sufficient sample and not an excess of time and handling	Even though the background cell barcodes are removed, this ambient signal exists across all cell barcodes
High mitochondrial percentage	Cells may be stressed during preparation or phenotype of those cells	If still good gene detection, may be okay	Beware of hard filters on percent mitochondria as they can exclude certain cell types
Low gene / UMI detection (after sufficient sequencing)	Cells may be stressed, may have less RNA, or something impeded RT	Gentler processing and handling, and full removal of chelators with washes	Additional sequencing will often increase detection, but not if complexity doesn't exist in cDNA library
Missing cell types	Either cells are lysed or removed during processing / enrichments or filtered out during data processing	Examine datapoints beyond the standard filters for expected markers, or consider a different sample prep method (nuclei?)	Some cell types may be particularly susceptible to loss (difficult to dissociate or fragile). Be aware of interpretations of data as cell survey.

You may have the best you can get, and the data can still be useful

Droplet based (10x Genomics) or plate-based (Smart-Seq) – when should one be considered over the other? What is the difference between high throughput single cell methods like droplet-barcoding (like 10x Genomics Chromium) and plate-based methods (like ‘Smart-Seq’)? Which is more sensitive? What is the comparative cost of each?



Droplet-based

(10x Genomics Chromium, DropSeq, DDSeq, Dolomite Nadia, etc.)

- High throughput
- ~40-50% or more of cells not captured
- Typically end-counting only
- “Cheaper” in that cost is ~\$0.25 / cell for the library

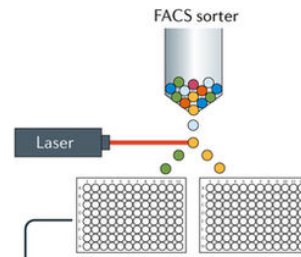


Plate or Integrated fluidic chip-based

(Fluidigm C1, Smart-Seq2, Takara SMARTer, Takara iCell8, Qiagen UPX, CellenONE, etc.)

- Sensitive; generally lower throughput
- Some allow full-length: isoform / variants
- Combinatorial indexing as an ultra high-throughput variation
- May require access to FACs or specialized equipment
- Smart-Seq2-like protocols more expensive (~\$20-60 / cell for libraries)



Microwell-based

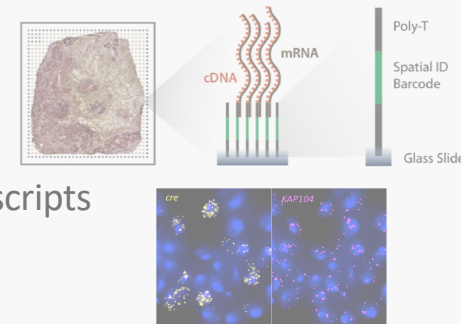
(BD Rhapsody, Celsee, MicroWell, etc.)

- High percent of cells captured
- Typically end-counting only
- Similar sensitivity to droplet-based?

Spatial gene expression

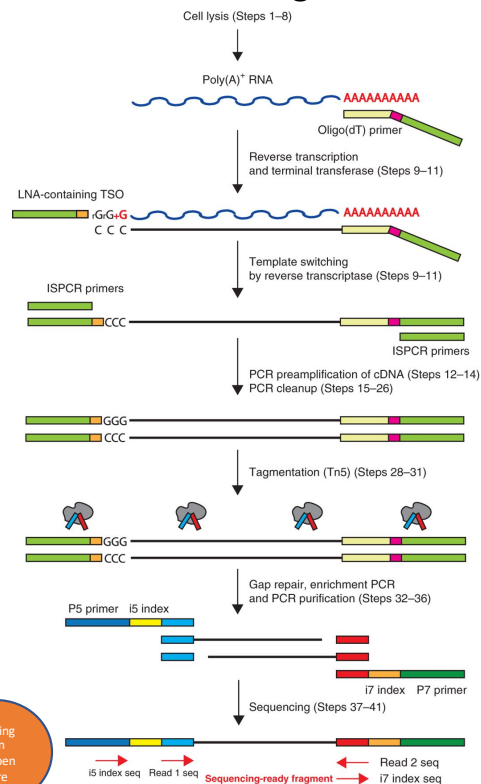
(10x Genomics Visium, NanoString DSP, Multiplex FISH, In Situ Seq etc.)

- Various approaches
- Inherent challenge of detecting all transcripts at true single cell resolution



Similarities and differences between full-length and end-counting scRNA-Seq library generation

Full Length



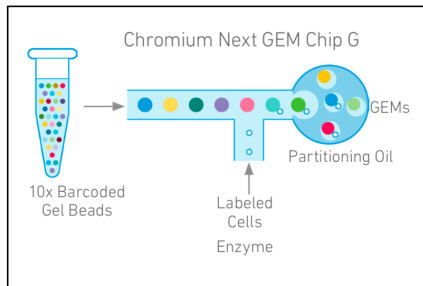
Full-length example - Smart-Seq2 Protocol

- Most protocols use reverse transcription (RT) to generate full length cDNA molecules
- End-counting methods add cell barcode at the RT step, allowing for early pooling and bulk processing of library
- Single cell per well full-length methods barcode at the stage when the library is fragmented for sequencing library prep – everything is kept in individual wells before that
- End counting methods enrich for transcript end fragments that contain cell barcodes
- Cell barcodes tell you which cell each molecule came from

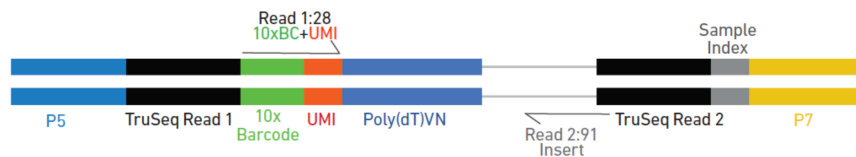
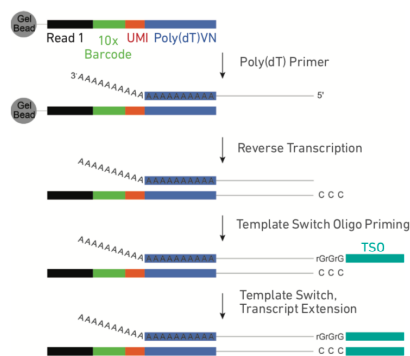
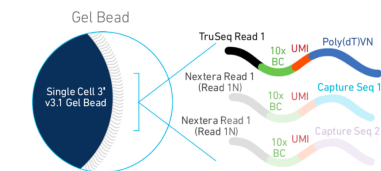
End-Counting



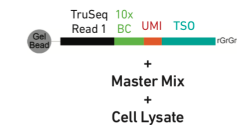
End-counting example - 10x Genomics 3' Protocol



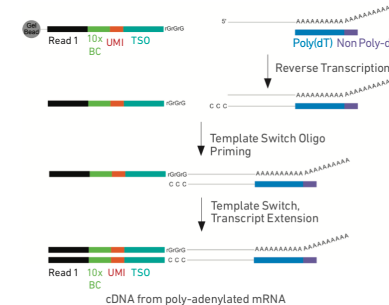
Difference between 3' and 5' gene expression profiling – which one should I use?



- Adding cell barcodes to 3' is generally more common
- Better sensitivity (arguable with improved 5')
- For feature barcode assays, need the Capture Sequence on the 3' bead (10x specific)

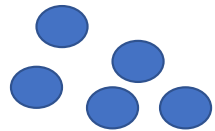


A.

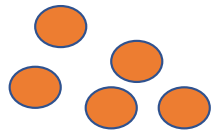


- Often used for immune profiling (gene expression + VDJ)
- Sensitivity improving; some prefer 5' data over 3'
- Adding cell barcodes to 5' end has some advantages

Measuring biology - avoid confounded design



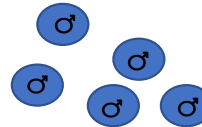
Sample1:
Condition1



Sample2:
Condition2

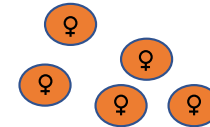
Biological comparison mixed with any other sample and capture factors. No way of checking or determining contribution.

Older
biological
age



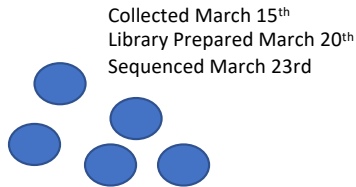
Sample1:
Condition1

Younger
biological
age

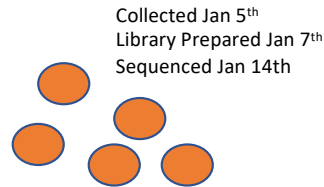


Sample2:
Condition2

Example1: Biological comparison may be confounded with age or gender of person / animal from which sample is sourced



Sample1:
Condition1



Sample2:
Condition2

Example2: Biological comparison may be confounded with technical variation related to sample capture or molecular biology

Rep1:
Condition1



Set1



Rep1:
Condition2

Rep2:
Condition1



Set2



Rep2:
Condition2

Rep3:
Condition1



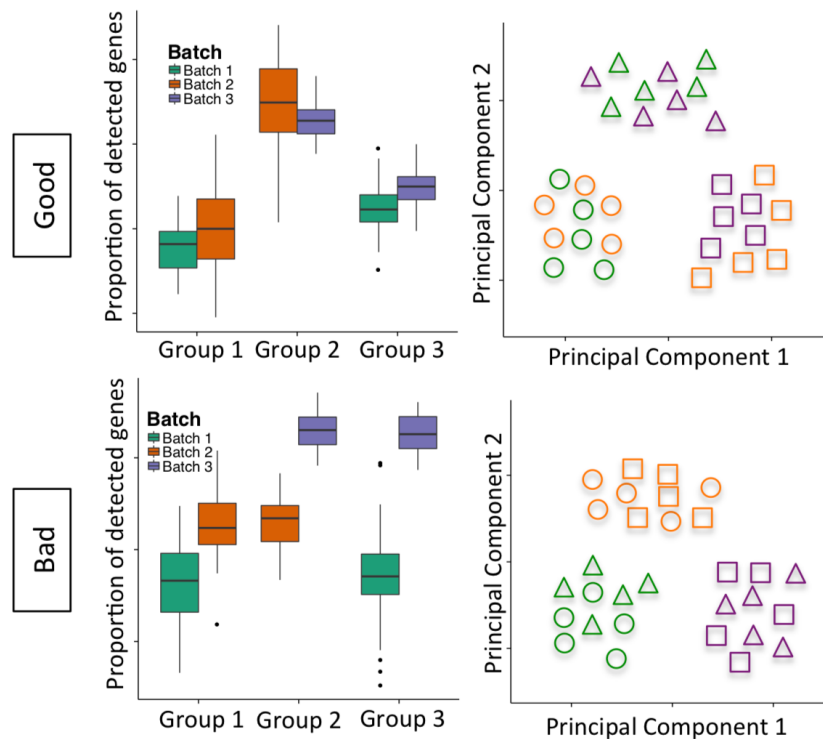
Set3



Rep3:
Condition3

Biological effect should be robust across replicates. Minimizing variation across comparison set is still good idea.

Confounded study design makes it difficult to separate biological from technical variation



From Hicks et al Biostatistics 2018

- Both biological and technical variation will exist in your dataset, and this becomes a problem when you have a weak biological signal or strong technical variation – how do you know which is dominant?
- Often there are practical considerations that impede a perfectly designed / balanced experiment
- It may help to have the person(s) who will be running the bioinformatic analysis involved in study design
- Don't make your design overly complicated in an effort to manage all variables
- Good design with replicates helps identify biological variation and prevents overcorrecting during technical batch handling, if needed (more on this later)

How many biological replicates?

Cost is often the biggest consideration in defining how many biological replicates to run. Experimental design is often a balance of cost and perfect design.

Isn't a single cell capture made up of thousands of biological replicates? Are replicate captures of the same prepared single cell suspension biological replicates?

Yes and no. The number of individual cells will increase statistical power, but you have to consider where confounding factors will still play a role

Two samples for capture and sequencing can cost \$5-\$7k

Pilot experiments are extremely helpful. If they can be included as the first biological replicate set, it is a win-win.

Consider sample multiplexing methods to increase biological replicates without a considerable increase in cost. *Some danger in multiplexing precious samples.*

Some practical questions to ask:

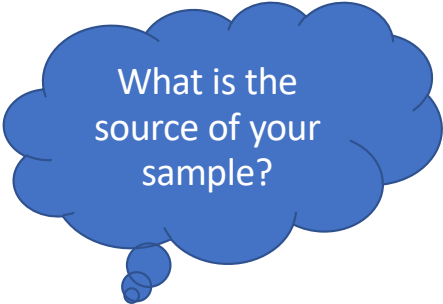
- Is your single cell sequencing going to be central to the study conclusions? What investment will be based off the data?
- Is the single cell data for validation / support of existing data?
- Do you have an independent method of validating?
- Where does an overly complicated design make high quality sample impractical to achieve?

Questions?

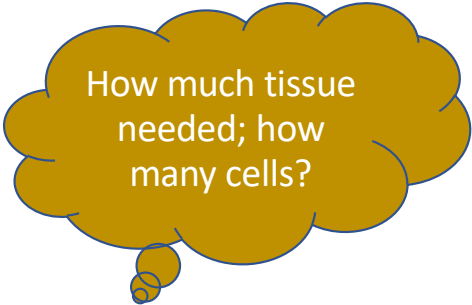
Sample preparation sets the
stage

Sample preparation may be the largest component to a successful single cell sequencing experiment.
Investing time and effort here is well worth it.


Some important questions to think about...



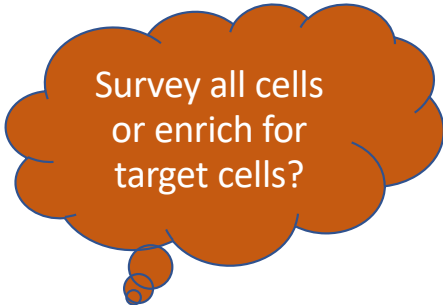
What is the source of your sample?



How much tissue needed; how many cells?

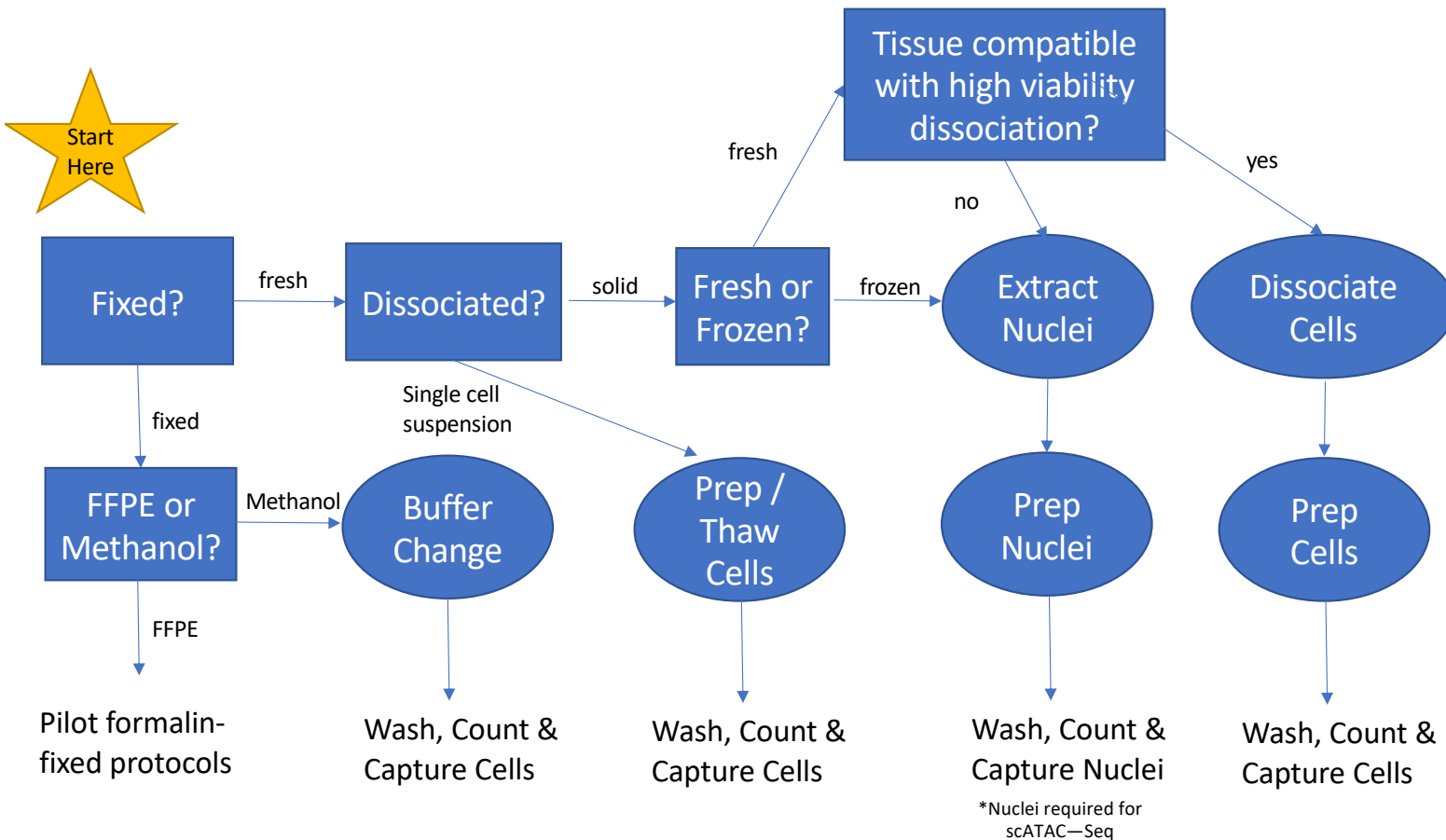


Are all cells viable after dissociation?



Survey all cells or enrich for target cells?

How to prep your sample for single cell sequencing



Other important notes:

- All methods require some optimization
- Does your sample have low viability cells or clumps of cells that need to be removed?
- Do you need to enrich for a cell type of interest?
- What effect does the sample preparation have on the thing you are trying to assay?

I targeted 5000 cells, why do I only have 1000 datapoints?

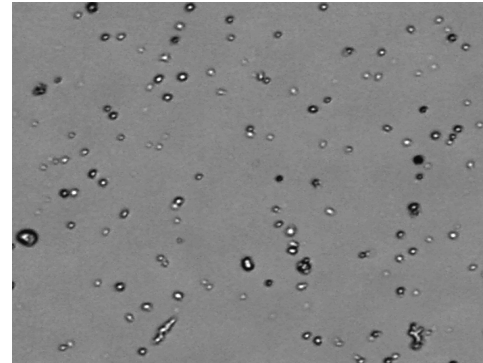
Large variability in number of datapoints in droplet-based single cell is common; expect a relatively large range.

Some factors are likely to play a strong role:

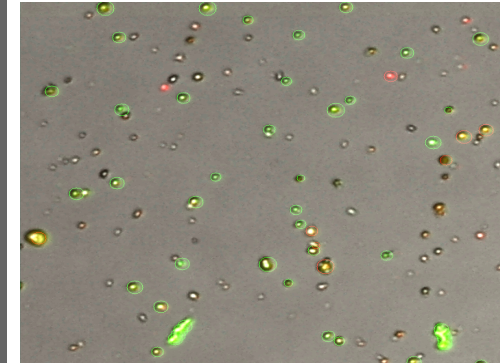
- Cell viability less than ideal (<90%)
- Inaccuracy of cell counts
 - Contaminating cells or debris add challenges
- Partial failure of the cell partitioning – be sure to check emulsions after each capture
- Primary cell samples tend to have more variability in cell size and RNA content

Some recommendations:

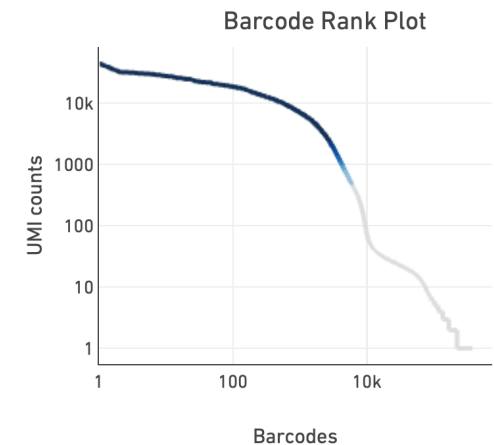
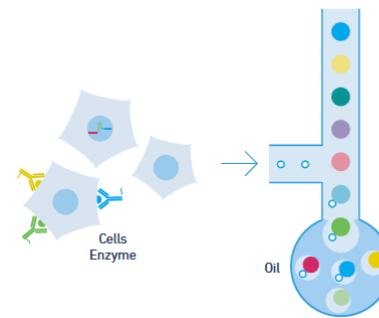
- Perform cell counting with a viability assay (AO-PI or similar may be more robust than Trypan)
- Factor in percentage of dead or contaminating cells – they may not generate datapoints, but they can contaminate signal



Brightfield Cell Count



Cell Count with Fluorescent Viability Assay (AO-PI)



Alternative sample prep methods when viable single cell dissociations are not practical

- **Single nuclei preparation**

- Fast extraction of nuclei from solid tissue; little dissociation-driven artifact
- Less RNA content than whole cell; higher pre-mRNA ratio
- Compatible with frozen tissue or difficult to dissociate tissue
- More difficult to QC sample; results assessed after sequencing

Matson et al 2018 JoVE

- **Transcriptional inhibition / cold-active proteases**

- Perform dissociation in transcriptionally-slowed environment
- Reduces dissociation-driven transcriptional artifact
- Additional control of dissociation process required

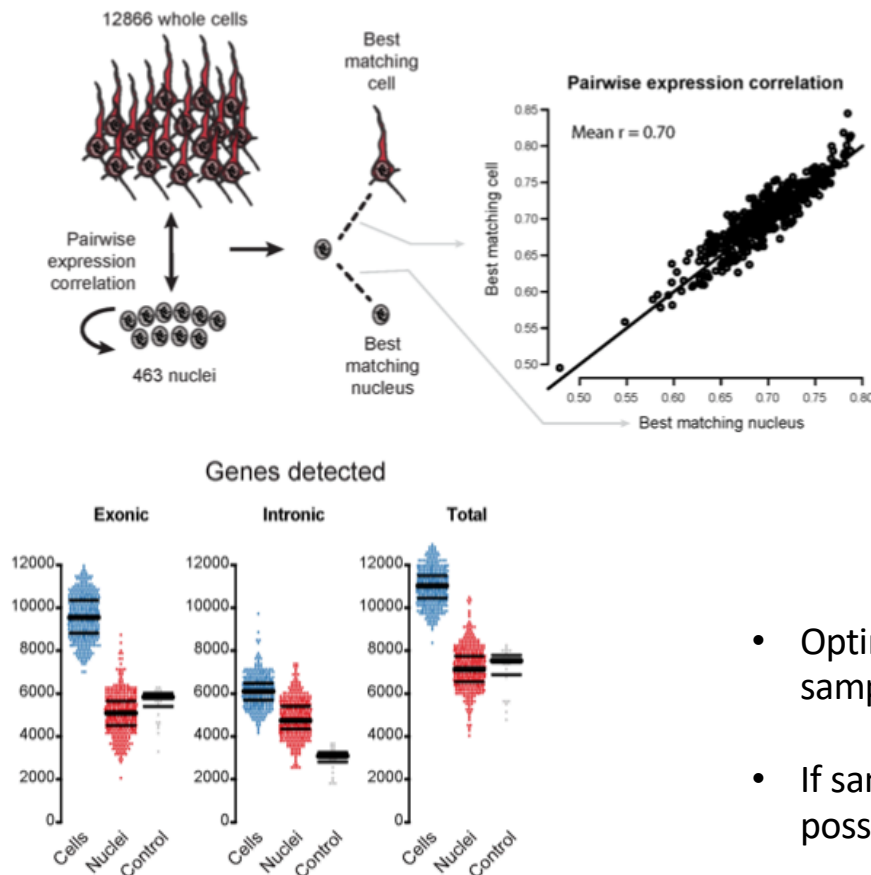
Wu et al 2017 Neuron

- **Cell fixation with Methanol**

- Dissociated cells still needed as input, but allows 'batching' of samples
- Preserves cells and transcript content for cold storage
- May not work for all cell types;

Chen et al 2018 J Transl Med

More about single nuclei RNA-Seq



Bakken et al. PLoS One, 2018

- Decent correlation between gene expression profiles from single cells and single nuclei
- Lower gene detection rate, with higher amount of intron retention (likely pre-mRNA)
- Good option for difficult to access tissue, etc.
- Required for single cell ATAC protocols, and will be required for combined snRNA-Seq/snATAC-Seq methods

- Optimization of robust nuclei extraction protocol not trivial, and sample viability doesn't work
- If sample is limited and/or precious, consider implications of possible sample loss

Some more detailed reading: Slyper et al Nature Medicine 2020

Sample prep summary

- Do the upfront work of establishing the best sample prep method – you'll save yourself many headaches (and overall cost) down the road
- Whole cell, high-viability single cell preparations may still be the best input for single cell RNA-Seq, but consider other methods if you have significant effects of dissociation or overall low viability
- Any manipulation or enrichment of cells may have an effect on the downstream data – know your process and keep it in mind when interpreting data
- Sometimes what you have is the best you can get – even a non-ideal dataset can still lead to great insight, but be prepared to validate before investing heavily in a potentially spurious result



June 18, 2020

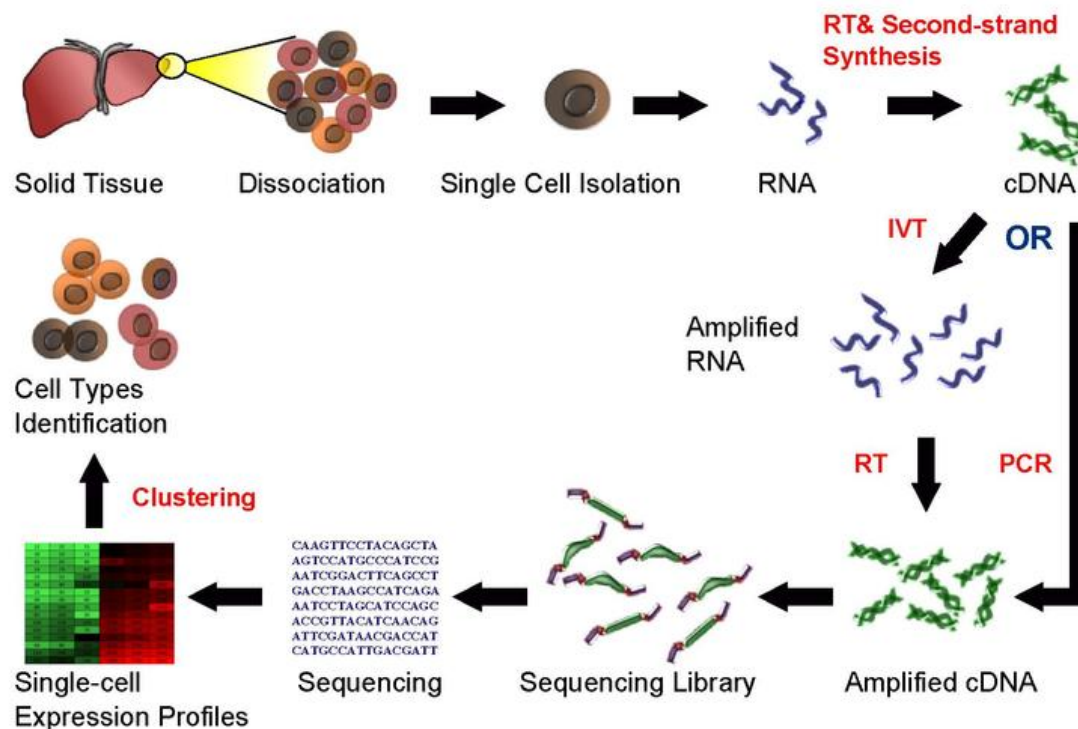
Sample Processing Considerations for Single Cell Sequencing – A Crucial Component of Experimental Design and Data Interpretation

Dr. Maria Hernandez, NIH/NCI/FNL

Platform and method define the
data type

Generalized workflow of generating single cell RNA-Seq data

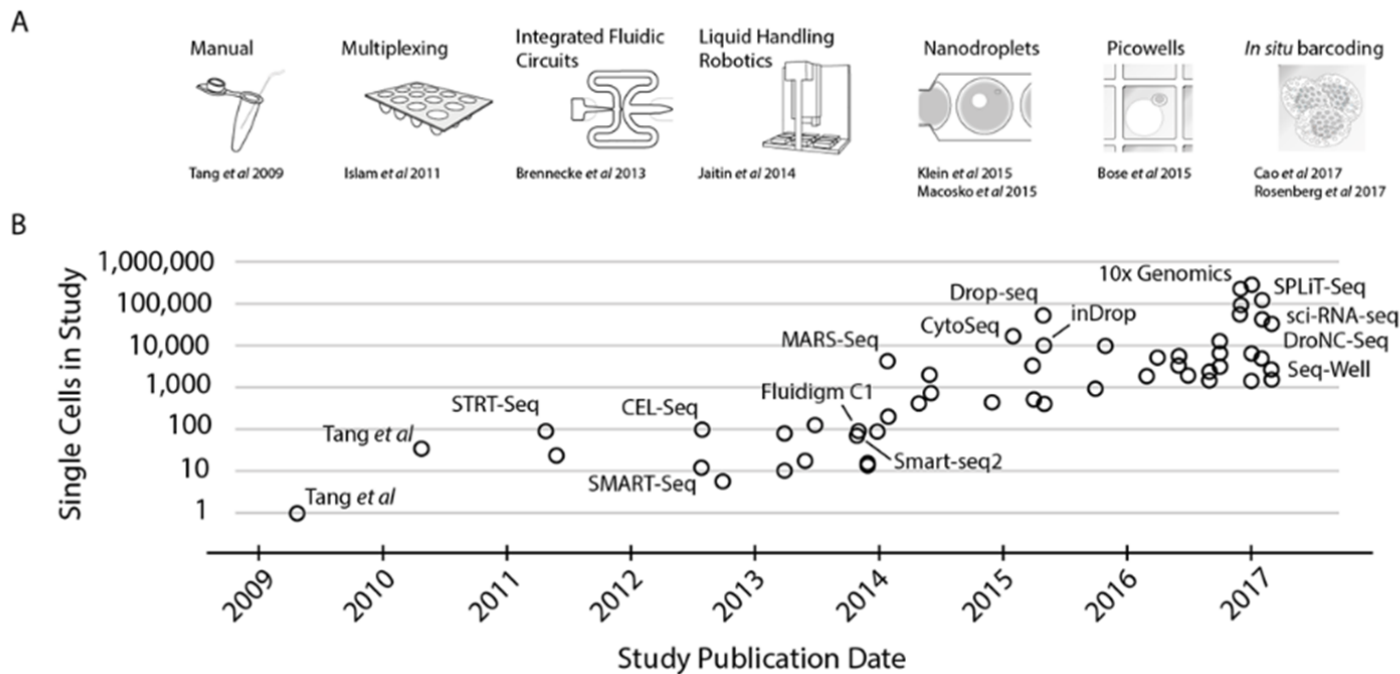
Single Cell RNA Sequencing Workflow



- Partition single cells
- Convert mRNA into cDNA
- Amplify cDNA
- Generate sequencing library
- Sequence
- Data analysis with identification of what transcripts are expressed by each cell profiled

<https://hemberg-lab.github.io/scRNA.seq.course/>

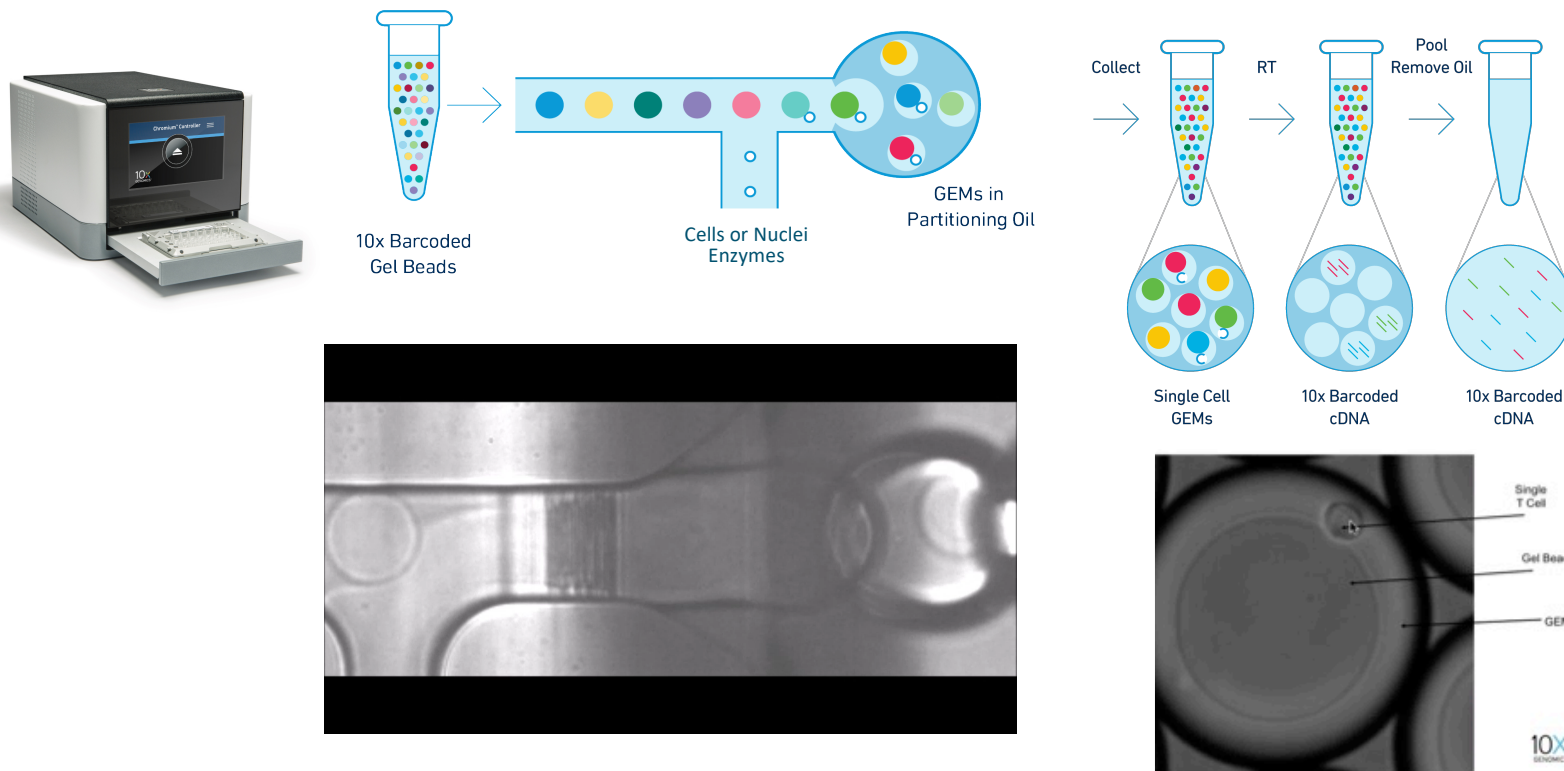
Single cell RNA-Seq has evolved quickly from lower throughput to higher throughput methods



Svensson et al. 2018

- First single cell whole transcriptome single cell RNA-Seq used manual picking of cells (2009)
- More widely adopted in 2012/2013 with Fluidigm C1 platform and SMARTer chemistry
- Huge increase in throughput with droplet based methods in 2015 (Drop-Seq / InDrops)
- Third generation of methods allow additional increase in throughput / decrease in cost (sciRNA-Seq / SPLiT-Seq / Seq-Well) ~2017/2018
- Spatial Profiling methods may be an additional frontier of 'single cell' data type

Droplet-based single cell sequencing has been dominant method for last few of years



- Fast partitioning of cells
- Early-stage barcoding of full-length cDNA molecules
- Enrichment of targets (such as VDJ) possible, but need to retain end with cell barcode
- Includes unique molecular identifiers

Sequencing depth investment
and strategy

Effect of sequencing depth on data sensitivity

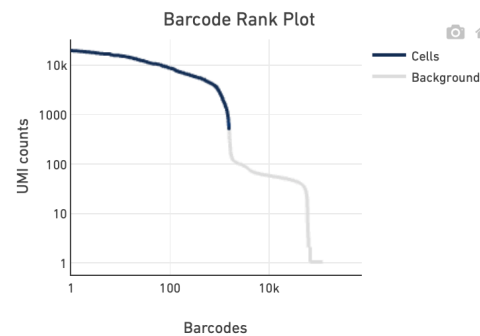
1,567
Estimated Number of Cells

143,978 **1,353**
Mean Reads per Cell Median Genes per Cell

Sequencing ?

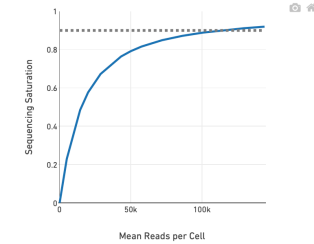
Number of Reads	225,613,243
Valid Barcodes	89.3%
Valid UMIs	99.8%
Sequencing Saturation	92.0%
Q30 Bases in Barcode	97.0%
Q30 Bases in RNA Read	91.1%
Q30 Bases in Sample Index	97.4%
Q30 Bases in UMI	96.3%

Cells ?

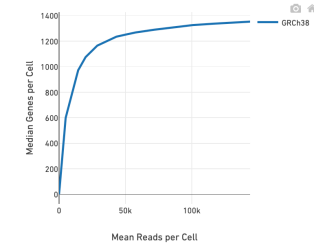


Estimated Number of Cells	1,567
Fraction Reads in Cells	71.5%
Mean Reads per Cell	143,978
Median Genes per Cell	1,353
Total Genes Detected	18,096
Median UMI Counts per Cell	3,815

Sequencing Saturation ?



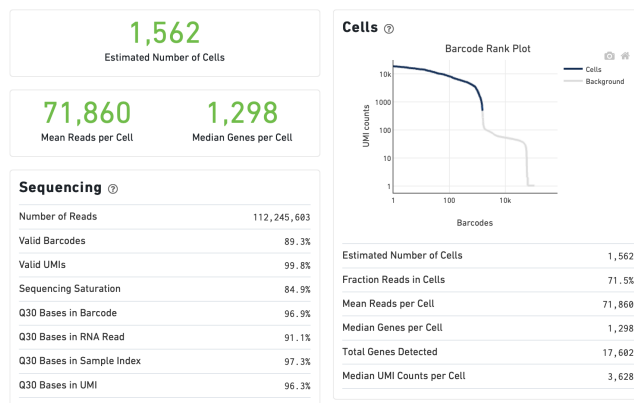
Median Genes per Cell ?



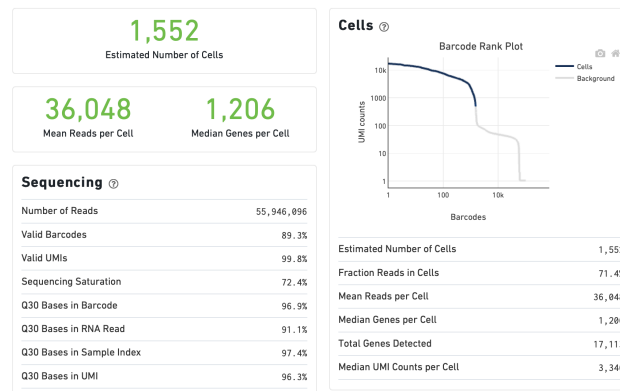
- More genes detected and better dynamic range with more UMIs at higher sequencing depth, but upper limit is complexity in original library
- Gains with higher sequencing diminish at higher depths
- Depth needed depends largely on goals of study

- Default target 50k reads / cell on average
- Evaluate sequencing saturation and gene detection projections
- Note that at low sequencing saturation the estimate can be quite inaccurate
- Other library type and goals may require different depth

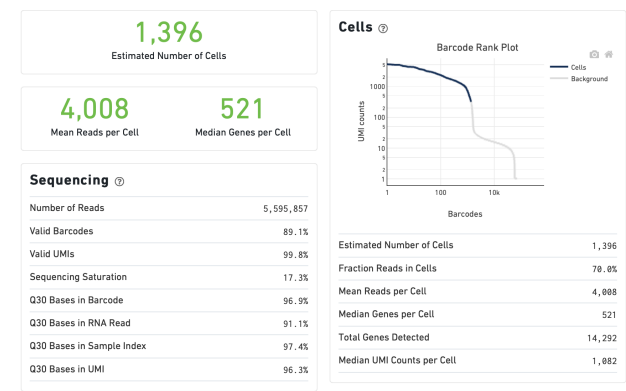
Effect of sequencing depth on data sensitivity



Estimate Cells: 1,562
 Reads per Cell: 71,860
 Genes per Cell: 1,298
 UMI per Cell: 3,628
 Seq Saturation: 94.9%



Estimate Cells: 1,552
 Reads per Cell: 36,048
 Genes per Cell: 1,206
 UMI per Cell: 3,346
 Seq Saturation: 72.4%



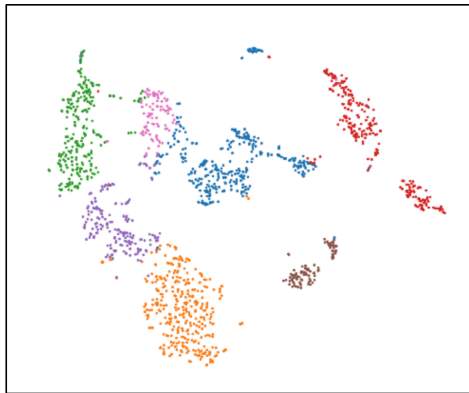
Estimate Cells: 1,396
 Reads per Cell: 4,008
 Genes per Cell: 521
 UMI per Cell: 1,082
 Seq Saturation: 17.3%

89% 71k/cell depth
 5% 71k/cell depth
 40% 71k/cell depth
 30% 71k/cell depth

- Many genes and UMIs still detected at moderate depth
- For relatively 'clean' datasets, very low sequencing can still reasonably estimate number of cells
- 20-fold increase in sequencing doesn't lead to 20-fold increase in sensitivity

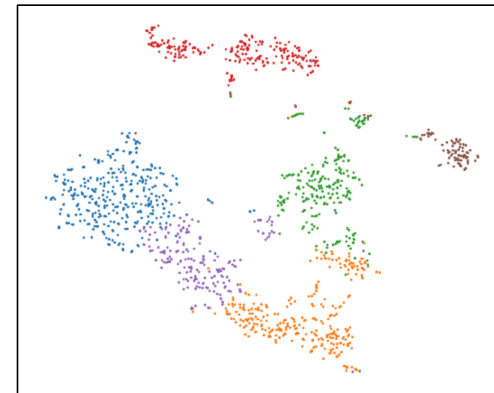
71k / cell
depth

Estimate Cells: 1,562
Reads per Cell: 71,860
Genes per Cell: 1,298
UMI per Cell: 3,628
Seq Saturation: 94.9%

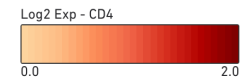
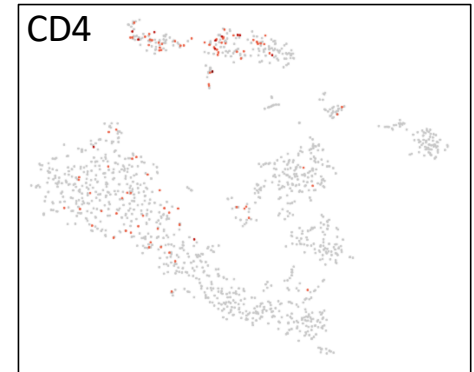
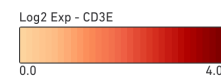
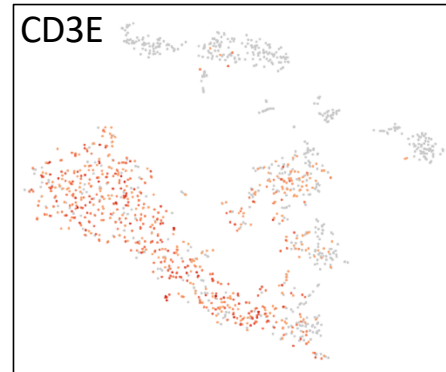
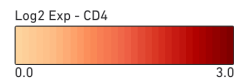
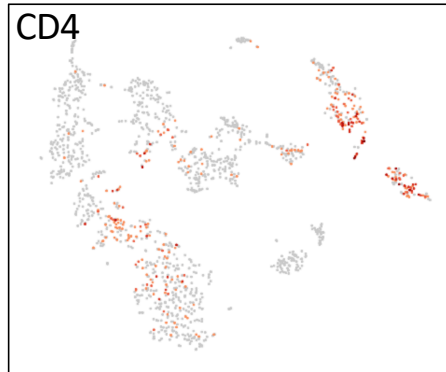
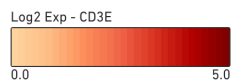
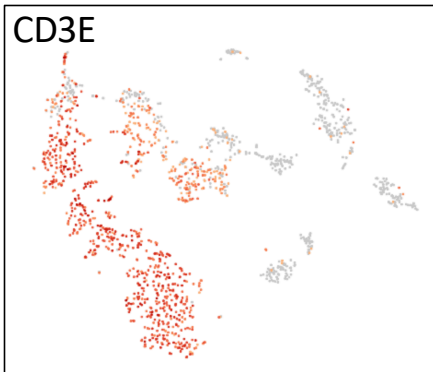


4k / cell
depth

Estimate Cells: 1,396
Reads per Cell: 4,008
Genes per Cell: 521
UMI per Cell: 1,082
Seq Saturation: 17.3%



*Similar
identification
of major cell
types; finer
separation in
higher depth*



*Range is half that
of higher depth*

Management of sequencing depth and cost is part of the plan and the process

Sequencing is cheaper at larger scale

	Cost	Number of paired reads	# cells covered at 50k reads / cell average	Cost per million reads
NextSeq 150-cycle High Output	~\$3,000	~400 M	~8,000	~\$7.50
NovaSeq 200-cycle S4	~\$27,000	~10 B	~200,000	~\$2.70

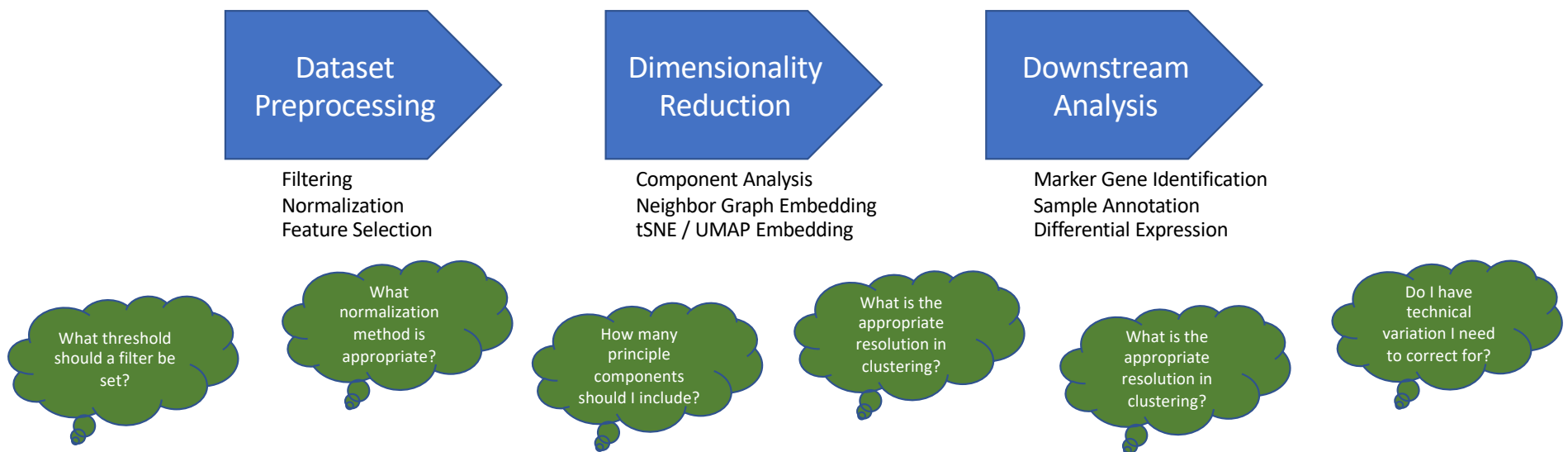
Although you could wait until a large set is completed and sequence together, there are several benefits of having information from initial sequencing as you collect samples:

- *Quality control check – it would be unfortunate to be many captures in before realizing something needs to be corrected*
- *Indication on whether experiment is work – low depth info can still be informative*
- *You can tailor your higher depth sequencing run to the number of cells represented in each capture*
- *You may also discover you might want more or less sequencing than original estimated*

Strategy: *Run lower depth sequencing first to evaluate data, and then follow-up with higher depth sequencing. Most trends will hold between the lower depth and higher depth sequencing. Initial sequencing data can be combined with the higher depth sequencing. Remember the unit price usually falls with increase in scale.*

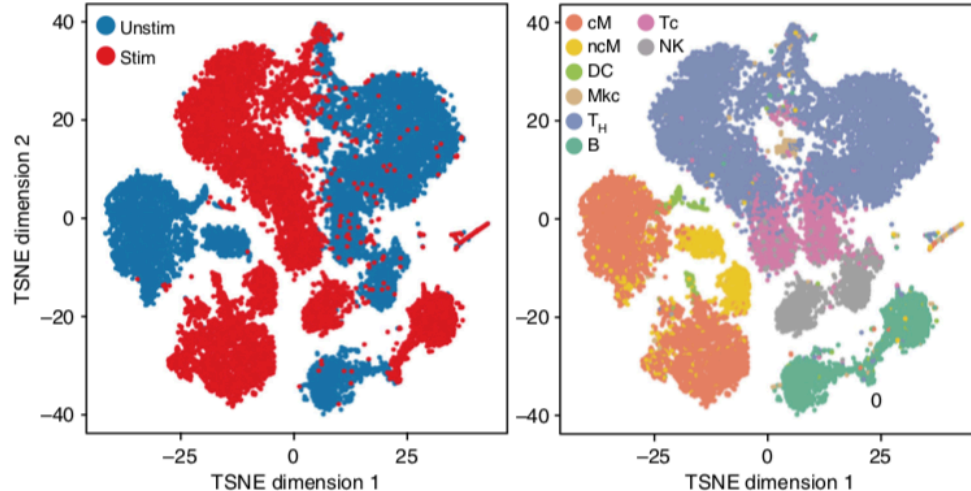
Analysis as an iterative process –
biology in focus

Generalized analysis workflows have shared components, but many parameter options exist



- The process is often iterative, but analysis with different parameters often shows similar underlying information
- The biology needs to inform the analysis - informatic analysis should go hand-in-hand with biology subject matter expertise and wet-lab processing
- Start with relaxed filter thresholds to ensure cell types not excluded and
- Avoid batch correction initially (and maybe completely) to ensure overcorrection does not occur
- Many new methods and tools (evaluate what is going to be helpful for end goal and be wary of oversimplification)

Some additional comments on batch correction, dataset alignment and overcorrection



Kang et al Nature Biotech 2017

Several powerful data alignment and batch correction methods exist, even allowing information transfer across modalities (scRNA-Seq – snATAC-Seq). Having some stable cell populations is typically required or aids their performance.

- How to compare cells across different timepoints or individuals if you don't know they are the same cell types?
- Is the differences observed biological or technical?
- Biological replicates can help provide an averaged signal for comparison
- Alignment or batch correction can adjust data to remove variation – important to only remove unwanted sources and not overcorrect

Some important questions to think about:

Are you performing dataset alignment for information transfer or batch correction?

When generating visualization or downstream analysis, is the original or corrected data used?

Do you lose important biological information after batch correction?